DOCUMENT RESUME

ED 087 815 TM 003 439

AUTHOR Kolakowski, Donald
TITLE Latent Trait Estimation: Theory vs. Practice.
PUB DATE 72
NOTE 11p.; Paper presented at American Psychological Association Symposium on the New Psychometrics

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS Item Analysis; Item Sampling; Measurement Techniques; Models; Prediction; *Psychological Testing; *Psychometrics; *Scoring; Standard Error of Measurement; Testing; Test Interpretation; *Test Reliability

ABSTRACT

Empirical results are presented as regards the implementation of a latent-trait psychometric model by means of conditional maximum likelihood estimation. Items are scored polychotomously into varying numbers of nominal categories and the test and item characteristic curves and information functions are examined. It is concluded that scoring items in four or more categories, as opposed to the usual dichotomous scoring, can increase information gain by a factor of two or more in the lower range of ability. Thus, the error of measurement is decreased to an extent equivalent to doubling the test length in this range. Alternatively, one can sample the range of ability in the target population with far rewer items. This latter property addresses itself directly to the empirical constraints on time and resources which are encountered in psychological testing. (Author)

APA Symposium on the New Psychometrics
Latent Trait Estimation: Theory vs. Practice

Donald Kolakowski
University of Connecticut

Empirical results are presented as regards the implementation of a latent-trait psychometric model by means of conditional maximum likelihood estimation. Items are scored polychotomously into varying numbers of nominal categories and the test and item characteristic curves and information functions are examined. It is concluded that scoring items in four or more categories, as opposed to the usual dichotomous scoring, can increase information gain by a factor of two or more in the lower range of ability. Thus, the error of measurement is decreased to an extent equivalent to doubling the test length in this range. Alternatively, one can sample the range of ability in the target population with far fewer items. This latter property addresses itself directly to the empirical constraints on time and resources which are encountered in psychological testing.

APA Symposium on the New Psychometrics
Latent Trait Estimation: Theory vs. Practice

Donald Kolakowski
University of Connecticut

The present study was an attempt to bring the theoretical advantages of latent trait estimation and multiple category scoring to bear upon a practical measurement problem.

These advantages include a continuous interval scale of measurement which is independent of the characteristics of the particular items employed, and an increase in precision due to the information recoverable from "wrong" responses; that is, responses which subjects exhibiting high levels of the trait in question would be very unlikely to choose. The problem in question is that of measuring verbal ability in a population of rural, disadvantaged youngsters.

The data consisted of the responses of 1,000 5th grade male subjects to a 10-item multiple choice reading subtest from the Survey Test of Educational Achievement. It required the testee to choose synonyms for words in context and to choose answers to questions about a short story. The test proved to be sufficiently difficult as to elicit "wrong" responses with better than chance frequency. Hence, scoring the items polychotomously held promise of recovering considerable information which would otherwise be discarded. However, there was no substantive or structural basis on which to rank the wrong alternatives. Therefore, the measurement model for nominal response categories proposed by Bock (1972) was adopted.

RESULTS

The data was analysed using the LOGOG program of Kolakowski and Bock (1972). Both multiple and dichotomous scoring schemes were investigated, using an empirical distribution of subjects into equal fractiles as well as under the assumption of a normal distribution of ability. The binary scoring scheme averaged about 56 sec. per program cycle while the multiple categories model averaged 2'23" on an IBM 360/65 computer. After six cycles, the item parameters were changing and the third significant digit under the normality condition, and at the second digit for the empirical prior. Limited resources prevented further computation. In addition to the item analysis, the average measurement error and the test reliability coefficient were computed by integrating over the trait distribution, again assuming normality. The results are presented in Table 1.

The uniformly significant values of Chi Square are a disappointment. In view of the rather low reliability coefficients, it would appear that the test was too hard. However, it is clear that multiple responses provide a much better fit than right/wrong scoring. In these data, the assumption of normality also tends to elevate the Chi Square. Thus it is with some caution that we point out the encouraging increase in average reliability of .12 for the multiple over the binary scoring. This corresponds to a decrease of .10 in the average Standard Error of Measurement. Of course, the decrease in error will not be constant over the entire range of ability. We can best investigate this, as a function of ability, in terms of its reciprocal, the information function.

Figure 1 is the graph of the test information as a function of ability. The test is most sensitive in the mid-range due to the fact that most test items are of intermediate difficulty. Nevertheless, increase in information for subjects of very low ability is more than doubled.

It is interesting to note the shift to the right of the Binary curve under the assumption of normality. This replicates the result obtained by Bock(1972) with very good-fitting data and therefore cannot be attributed to the present lack of fit. On the other hand, a convergence problem occurred in the binary analysis which necessitated the elimination of a group of the lowest subjects while estimating the item parameters. Thus, the question of whether dichotomous scoring can always be expected to yield more information than multiple scoring for high trait levels remains indeterminate.

Focusing now on individual items, Figures 2 - 5 depict the information and operating characteristics of the item with the best fit, number 9, and with the worst fit, no. 6, in all four analyses. As with the test as a whole, assuming a normal prior shifted the mode of the binary information curve out from under the curve for multiple categories. Otherwise there is very little to choose between them. The monotonically increasing "best" answer is characterized by the largest slope estimate; the monotonically decreasing curve, the smallest. Both items increase their precision of measurement below the median by a factor of two or more, and are roughly equivalent to the binary scoring in the high range. However, under the normal prior, both binary information curves have a maximum about equal in magnitude to those for the multiple scoring. In the case of the empirical prior, these maxims are on the order of one-half that for the multiple case. Thus the difference in the Chi Square statistics for these two items does not indicate any gross abnormality in the behavior or magnitude of the characteristics and parameters of item 6, as compared with item 9. It does indicate greater deviations of the data points from their expected values, but such deviations often occur in only one or two of the operating curves for a particular item. Therefore, we are well advised to look beyond global statistics such as the total Chi Square. They may be too sensitive to deviations which have no substantive meaning and which introduce no systematic bias.

In conclusion, we can fairly say that scoring test items in four or five categories can decrease the error of measurement to an extent equivalent to doubling the test length for a certain range of ability or, alternatively, can sample the range of ability in the target population with far fewer items. This property addresses itself directly to the empirical constraints on time and resources which are encountered in psychological testing. While the model did not appear to fit the present data in terms of the Chi Square statistics, we have seen that substantive interpretation of the behavior of item alternatives did not seem to be impaired.

TABLE 1

Goodness of Fit

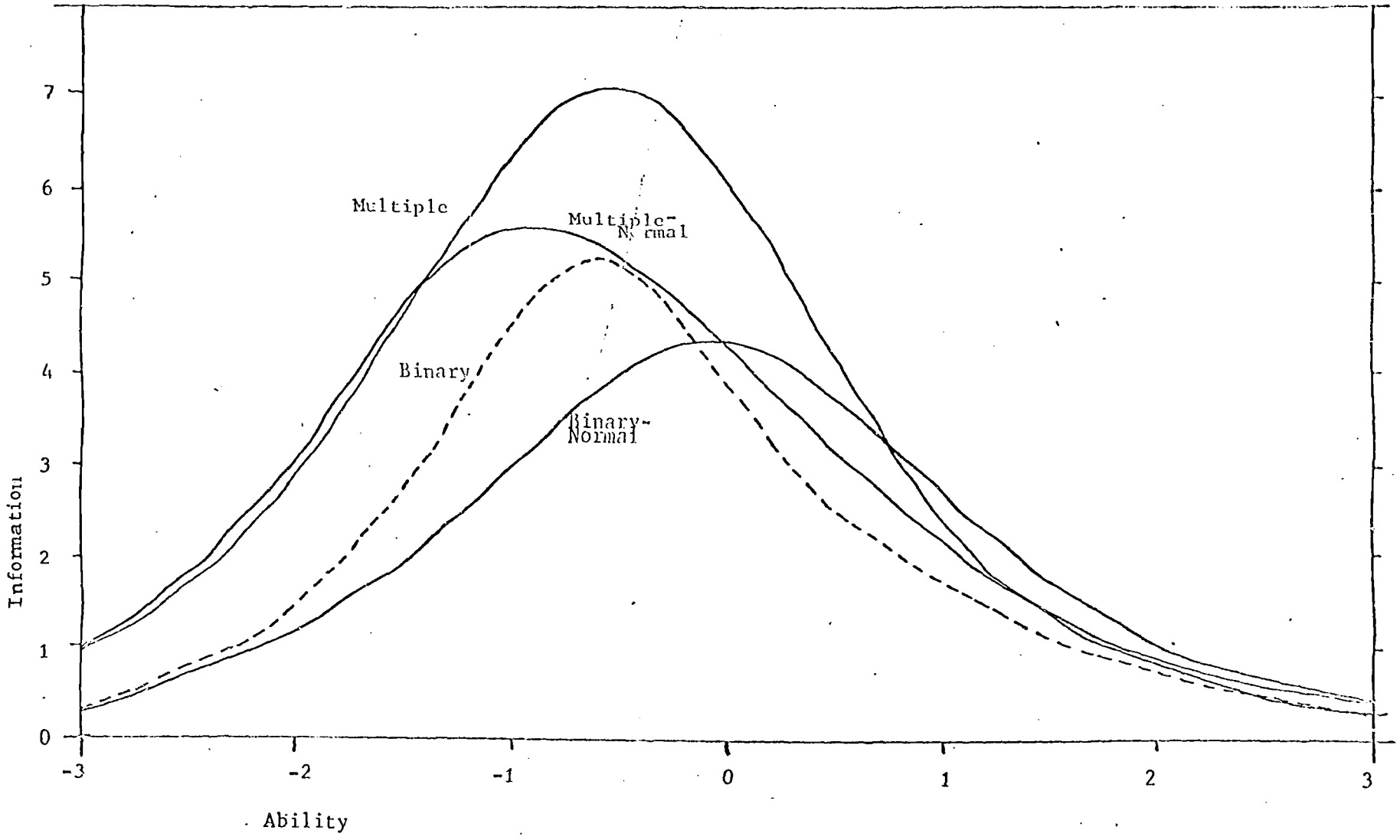
Scoring scheme	Distribution	Chi Sq./d.f.	r_{xx}	S.E.M.
Multiple	empirical	1.13 (2% level)	.70	.55
Binary	empirical	1.71	.58	.65
Multiple	normal	1.90	.66	.58
Binary	normal	2.67	.62	.62

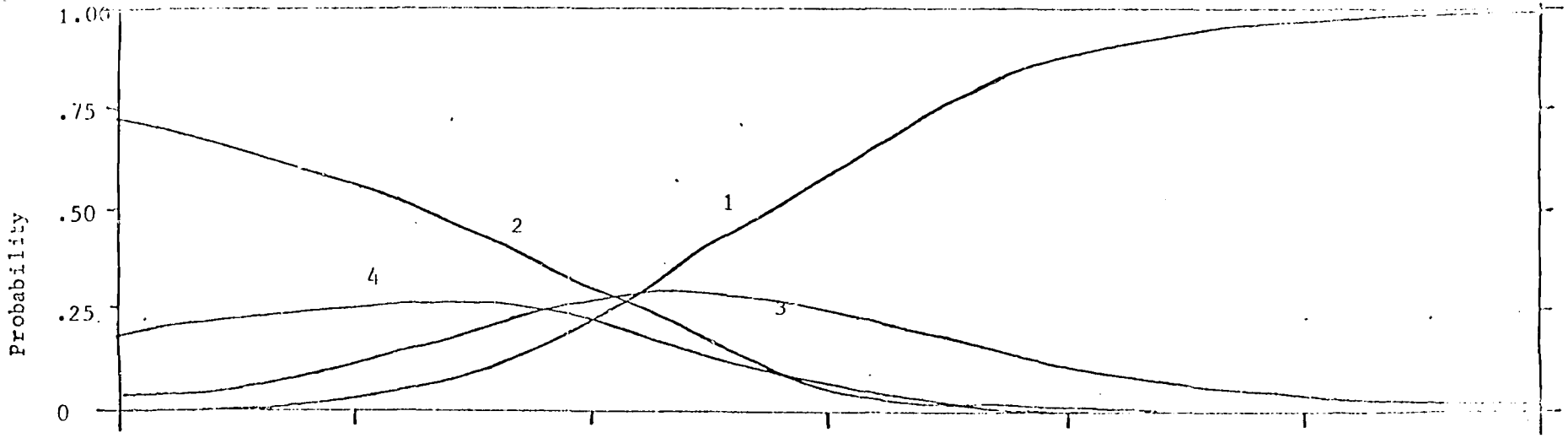
TABLE 2

Item Calibration

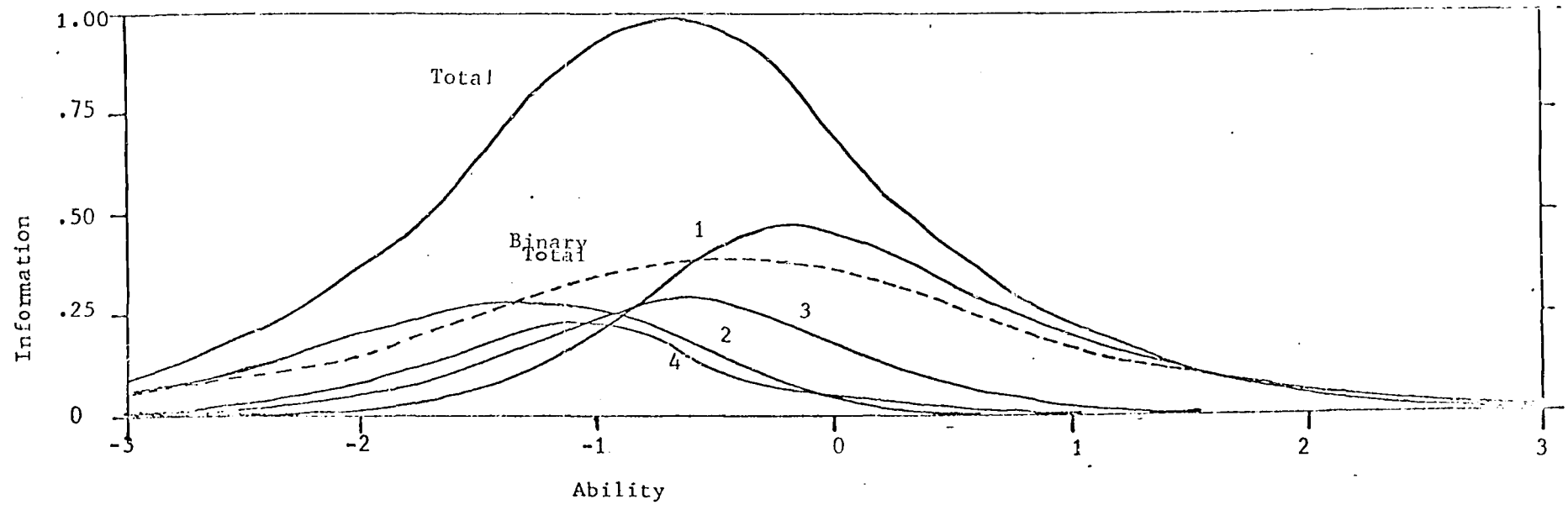
Response	Poor Item (6)				Good Item (9)			
	empirical prior		normal prior		empirical prior		normal prior	
	intercept	slope	intercept	slope	intercept	slope	intercept	slope
1	.187	1.658	.161	1.42	1.32	1.45	1.22	1.08
2	-.037	-.465	.0259	-.309	-0.914	-1.10	-.867	-.951
3	-.573	-1.305	-.613	-1.29	.367	.291	.296	.227
4	.422	.112	.425	.176	-.775	-.636	-.654	-.360
	$X^2(54)=79$		$X^2(24)=104$		$X^2(54)=32$		$X^2(24)=24$	
	Binary: $X^2(17)=39$		Binary: $X^2(8)=45$		Binary: $X^2(17)=24$		Binary: $X^2(8)=9$	

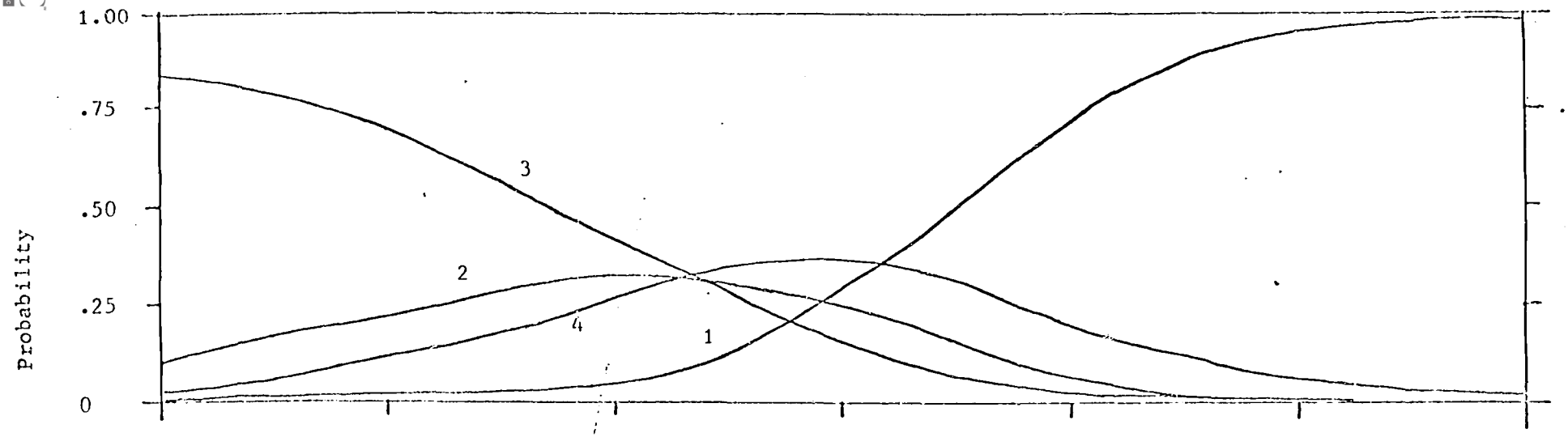
TEST INFORMATION FUNCTIONS



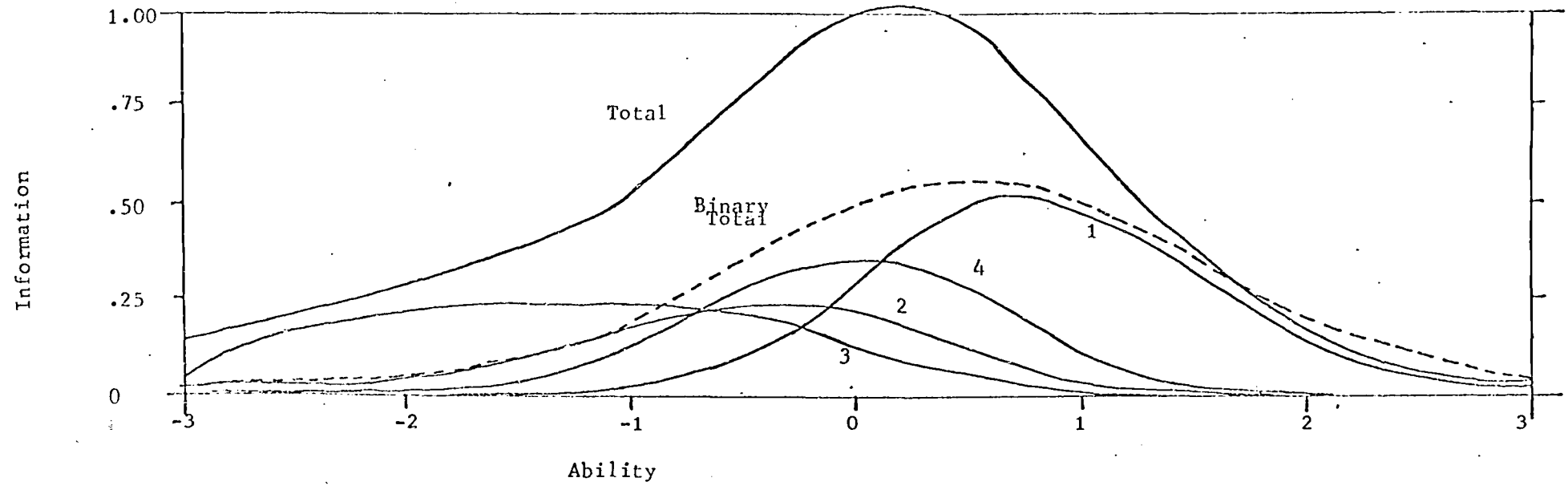

1.00
.75
.50
.25
0
Probability
1
2
3
4

ITEM 9


1.00
.75
.50
.25
0
Information
Total
Binary
Total
1
2
3
4
-3
-2
-1
0
1
2
3
Ability


1.00
.75
.50
.25
0
Probability
3
2
4
1

ITEM 6


1.00
.75
.50
.25
0
Information
Total
Binary
Total
1
4
2
3
-3
-2
-1
0
1
2
3
Ability